

Active prediction in graphical models [★]

Satyaki Mahalanabis and Daniel Štefankovič

Department of Computer Science
University of Rochester
{smahalan, stefanko}@cs.rochester.edu

Abstract. Given a graph with an unknown labeling of its nodes, we consider the problem of choosing a subset of nodes to observe so as to minimize the prediction error for labels of the remaining unobserved nodes. While this problem has been studied before, the novelty of our setting is twofold. First, we assume that the node labels are random variables having a known joint distribution, i. e., they constitute a “graphical model”. Second, we consider adaptive node selection strategies, i. e., the node we choose to observe at any given time may depend on the labels we have observed so far. We first prove a NP-hardness result for finding the optimal set of nodes for general graphical models (either adaptively or non-adaptively). Then we show that for ferromagnetic Ising models on the 1-D chain graph, adaptive strategies outperform the non-adaptive ones (we show this by proving matching upper and lower bounds for adaptive and non-adaptive strategies).

1 Introduction

Given an undirected graph with an unknown labeling of its nodes (using a fixed set of labels) and the ability to observe the labels of a subset of the nodes, consider the problem of predicting the labels of the unobserved nodes. This is a widely studied problem in machine learning, where the focus until now has largely been on designing the prediction algorithm. Recently however, the problem of selecting the set of vertices to observe (i.e. active prediction) has been receiving attention [5], and this is what our work tries to address. We are interested in designing adaptive strategies for the selection problem, i. e., at any point in time, the next node we choose to observe can depend on the labels of nodes that have already been selected and (hence) revealed. While previous research [7, 5] has mostly focused on the worst case prediction error of the selection strategy (i. e., assume adversarial labeling of all nodes), we analyze the expected error under the assumption that the labels are random variables with a known joint distribution. In other words we consider graphical models (such as the Ising model) with known parameters. We assume that the selection strategy is provided a budget of the number of nodes it can observe, and we study the expected number of mispredictions as a function of this budget. We

[★] Research supported, in part, by NSF grant CCF-0910584.

are particularly interested in the question of whether there exist very simple yet optimal node selection strategies.

The example we study is the (ferromagnetic) Ising model (one of the simplest graphical models) on a chain graph. To further simplify the problem we consider a continuous version of the Ising model, that is, the limit as the number of vertices on the path goes to infinity. We show that some simple adaptive selection strategies have optimal error (up to a constant factor), and have asymptotically smaller number of mispredictions than non-adaptive strategies. We also prove the hardness of selecting the optimal set of nodes (either adaptively or non-adaptively) for graphical models on general graphs.

1.1 Related work

In [7] authors consider the problem of non-adaptive selection for minimizing the worst case prediction error for the case of binary labels. They give an upper bound on the error of the min-cut predictor, for a given set of observed nodes S , in terms of the cut size (that is, the number of edges whose endpoints have different labels) of the unknown labeling and a graph cut function $\Psi(S)$ (which depends on the underlying graph). The problem of choosing the observed nodes then reduces to that of minimizing $\Psi(S)$ over sets S of a given (budget) size, for which they suggest a heuristic strategy (the goal of that paper is not to provide provable guarantees for the heuristic). Authors of [5] give a non-adaptive selection strategy in a similar adversarial setting for the special case of trees such that the min-cut predictor’s number of mispredictions is at most a constant factor worse than any (non-adaptive) prediction strategy (analyzed in the worst case for labelings that have bounded cut size). We point out that the algorithm of [5] does not need to know the cut size constraint while in our case, the expected cut-size is known and depends on the model parameters. Further, [5] provides a lower bound on the prediction error for general graphs as well. Unlike [7, 5], [1] considers an adaptive version of the selection problem for (certain classes of) general graphs where the cut size induced by the labeling is known. Their goal, unlike ours, is to recover the correct label of every node using as few observations as possible. They claim to give a strategy whose required number of labels matches the information-theoretic lower bound (ignoring constant factors).

Finally note that our goal here, namely, the active selection of nodes to observe, is different from that of, e. g., [4] where the nodes to be labeled are chosen by an adversary, or from that of semi-supervised settings, e. g., [2, 11] where the query nodes are chosen randomly. Also, we aim to minimize the expected number of mispredictions, as opposed to inferring the most likely labeling (i. e., the “minimum energy configuration”) of the unobserved nodes which is what most applications in computer vision tend to focus on, e. g., [3] (see also [9]).

2 The model

Graphical models are a powerful method of representing complex joint distributions exploiting the independence between components to reduce the number of parameters. In an undirected graphical model a distribution of (X_1, \dots, X_n) is described in terms of factors—functions ϕ_1, \dots, ϕ_m from subsets of variables to non-negative real numbers. The joint distribution of (X_1, \dots, X_n) is then

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{j=1}^m \phi_j(X_{S_j}),$$

where X_S denotes the vector of variables whose indices are in S and Z is the normalizing factor (called the partition function).

We are going to assume that the graphical model is known to us, a sample has been taken from it, and our goal is to predict the sample. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample from the graphical model. A prediction strategy A , after querying some of the variables outputs a vector $A(\mathbf{X}) = \mathbf{Y} = (Y_1, \dots, Y_n)$. The *average error* of the prediction is

$$\text{av-err}(\mathbf{X}, \mathbf{Y}) = \frac{|\{i \mid X_i \neq Y_i\}|}{n}.$$

The *expected average error* of the prediction strategy A is

$$\text{err}_A = \mathbf{E}[\text{av-err}(\mathbf{X}, A(\mathbf{X}))], \tag{1}$$

where the expectation is taken over the random choice of \mathbf{X} sampled from the model. Of course, if the strategy is allowed to query all variables in the model then it can achieve zero expected average error. The interesting question is—how well can we infer the values of all the variables if we can query only a few of them? Thus we are going to study the tradeoff between the number of queries and the expected average error.

The simplest type of prediction strategy is *non-adaptive*, that is, one that queries a subset of the variables $U \subseteq \{1, \dots, n\}$ and then makes a prediction. The prediction that minimizes the expected average error is to label each variable by the most likely label, given the observations, that is,

$$Y_i = \arg \max_a P(X_i = a \mid X_U = \hat{X}_U), \tag{2}$$

where \hat{X}_U are the values observed.

We are going to focus on *adaptive* strategies where the choice of variables queried can depend on the values of previously queried variables. Such a strategy can be described (in principle, here we ignore the fact that the description has size exponential in the number of queries) by a decision tree, where each internal vertex of the tree corresponds to a variable queried, edges correspond to values,

and each leaf gives a prediction for the unseen variables. If the set of variables queried on a path is U and the values observed are \hat{X}_U , then, again, the optimal prediction at that leaf is given by (2).

Question 1 *Given a graphical model, what is the best (adaptive) strategy with budget B ?*

We explore the computational aspect of Question 1—can the strategy be found in polynomial time? Unfortunately the answer is no, given a graphical model, figuring out the optimal non-adaptive (or adaptive) strategy with a given budget is NP-hard, this can be shown by reducing from the following problem (whose NP-hardness follows from CLIQUE):

DENSE B -SUBGRAPH PROBLEM:

INSTANCE: undirected graph H , budget B , threshold T .

QUESTION: does there exist $S \subseteq V(H)$, $|S| = B$ such that $|E(H) \cap \binom{S}{2}| \geq T$?

Proposition 1. *Computing the non-adaptive (or adaptive) strategy with budget B that minimizes the expected average error is NP-hard.*

The proof is deferred to Section 4.

One can still hope for approximately optimal strategy and/or to identify a family of models where the problem is tractable. (It is likely that some assumptions about the family might be needed—this is supported by the fact that we don’t know a good approximation algorithm for the dense B -subgraph problem [6]).

Question 2 *Given a graphical model, can one find approximately optimal (adaptive) strategy with budget B ?*

In the next section we explore the model on the Ising model and compare the adaptive and non-adaptive strategies.

3 Ising model in 1-d

The *Ising model* [8] is perhaps the simplest graphical model. We have an undirected graph $G = (V, E)$, the variables correspond to the vertices and the factors correspond to edges. Each variable has two possible values -1 and $+1$. We are going to consider the ferromagnetic Ising model with no external field, for which the *energy* of a configuration $\sigma : V \rightarrow \{-1, +1\}$ is given by

$$H(\sigma) = - \sum_{uv \in E} \sigma_u \sigma_v. \quad (3)$$

The distribution described by the model (called Gibbs distribution) is

$$P(\sigma) = \exp(-\beta E) / Z(\beta),$$

where $\beta > 0$ is a parameter (called *inverse temperature*) and $Z(\beta)$ is a normalizing constant (called *partition function*) that makes $P(\sigma)$ into a distribution.

Assume that the underlying graph is a path with n vertices. There are $2^{\binom{n-1}{k}}$ configurations that have k edges whose endpoints are labeled by different spins (the remaining $n-1-k$ edges have endpoints whose spins agree). Thus the partition function is

$$Z(\beta) = \sum_{k=0}^{n-1} 2^{\binom{n-1}{k}} \exp(\beta(n-1-2k)) = 2^n \cosh(\beta)^{n-1}.$$

Let

$$\begin{aligned} Z_A(\beta) &= \sum_{k=0, k \text{ even}}^{n-1} 2^{\binom{n-1}{k}} \exp(\beta(n-1-2k)) \\ &= 2^{n-1} (\cosh(\beta)^{n-1} + \sinh(\beta)^{n-1}), \end{aligned}$$

and

$$\begin{aligned} Z_D(\beta) &= \sum_{k=0, k \text{ odd}}^{n-1} 2^{\binom{n-1}{k}} \exp(\beta(n-1-2k)) \\ &= 2^{n-1} (\cosh(\beta)^{n-1} - \sinh(\beta)^{n-1}). \end{aligned}$$

The probability that the endpoints of the path have the same spin is given by

$$P_{A,n}(\beta) := \frac{Z_A(\beta)}{Z(\beta)} = \frac{1 + \tanh(\beta)^{n-1}}{2}. \quad (4)$$

The probability that the endpoints of the path have a different spin is given by

$$P_{D,n}(\beta) := \frac{Z_D(\beta)}{Z(\beta)} = \frac{1 - \tanh(\beta)^{n-1}}{2}. \quad (5)$$

Suppose that both endpoints are labelled by +1. Then the probability that the i -th vertex is labelled by +1 is given by

$$P_{AA,n}(\beta, i) := \frac{P_{A,i} P_{A,n+1-i}}{P_{A,i} P_{A,n+1-i} + P_{D,i} P_{D,n+1-i}} = \frac{1}{2} + \frac{\lambda^{i-1} + \lambda^{n-i}}{2(1 + \lambda^{n-1})}, \quad (6)$$

where $\lambda = \tanh(\beta)$ (note that $\lambda \in (0, 1)$). Suppose that left endpoint is labelled by +1 and the right endpoint is labelled -1. Then the probability that the i -th vertex is labelled by +1 is given by

$$P_{AD,n}(\beta, i) := \frac{P_{A,i} P_{D,n+1-i}}{P_{A,i} P_{D,n+1-i} + P_{D,i} P_{A,n+1-i}} = \frac{1}{2} + \frac{\lambda^{i-1} - \lambda^{n-i}}{2(1 - \lambda^{n-1})}. \quad (7)$$

If a strategy queried the endpoints and they came out labelled +1 then the strategy should predict label +1 for all vertices (this follows from (2) and the

fact that expression (6) is greater than $1/2$ for all i). The expected average error (conditioned on both endpoints being labelled $+1$) is then

$$E_{A,n}(\beta) := \frac{1}{2} - \frac{1}{n} \sum_{i=1}^n \frac{\lambda^{i-1} + \lambda^{n-i}}{2(1 + \lambda^{n-1})} = \frac{1}{2} - \frac{1}{n} \frac{1 - \lambda^n}{(1 - \lambda)(1 + \lambda^{n-1})}. \quad (8)$$

If an strategy queried the endpoints and the leftmost one is labelled $+1$ and the rightmost one is labelled -1 then the strategy should predict vertices in the left half to be $+1$ and the vertices in the right half to be -1 (again, this follows from (2) and the fact that (7) if at least $1/2$ for $i \leq n - i$ and at most $1/2$ if $i \geq n - i$). The expected average error (conditioned on both endpoints being labelled by disagreeing labels) is then

$$E_{D,n}(\beta) := \frac{1}{2} - \frac{1}{n} \left(\sum_{i=1}^{\lfloor n/2 \rfloor} \frac{\lambda^{i-1} - \lambda^{n-i}}{2(1 - \lambda^{n-1})} - \sum_{i=\lfloor n/2 \rfloor + 1}^n \frac{\lambda^{i-1} - \lambda^{n-i}}{2(1 - \lambda^{n-1})} \right) = \frac{1}{2} - \frac{1}{n} \frac{(1 - \lambda^{\lfloor n/2 \rfloor})(1 - \lambda^{\lceil n/2 \rceil})}{(1 - \lambda)(1 - \lambda^{n-1})}. \quad (9)$$

Remark 1. Suppose that we used the labels to predict the most likely configuration and used that to define some error measure. Now if, say, the endpoints have disagreeing labels then all configurations with one flip (that can occur on any of the edges) all have the same likelihood (and are the most likely configurations). However if we changed the interaction energy slightly on just one of the edges (by changing the constant 1 of $\sigma_u \sigma_v$ in (3) to $1 - \varepsilon$) then the most likely configuration would have flip on that edge—a drastic change produced by a small change in the model; from this it seems unlikely that a robust error measure can be defined this way.

The behavior of the model will be easier to understand in the continuous limit. Letting $\ell = -\frac{n-1}{2} \ln \lambda$ and $x = -\frac{i-1}{2} \ln \lambda$ in equations (6) and (7) we obtain

$$P'_{AA}(\ell, x) := \frac{1}{2} \left(1 + \frac{\cosh(\ell - 2x)}{\cosh(\ell)} \right) \quad (10)$$

and

$$P'_{AD}(\ell, x) := \frac{1}{2} \left(1 + \frac{\sinh(\ell - 2x)}{\sinh(\ell)} \right). \quad (11)$$

As $n \rightarrow \infty$ equations (8) and (9) converge to

$$E'_A(\ell) := \frac{1}{2\ell} (\ell - \tanh(\ell)) \quad \text{and} \quad E'_D(\ell) := \frac{1}{2\ell} (\ell - \tanh(\ell/2)). \quad (12)$$

Thus as $n \rightarrow \infty$ and β scales appropriately (so that ℓ stays fixed) expected average error of the path (with no labels except at endpoints) goes to $E'_A(\ell)$ and $E'_D(\ell)$, depending on whether the labels of endpoints agree or disagree. To make

keeping track of contributions of subintervals to the expected average error easier (after we make queries) we will sometimes use un-normalized expected average error which is just the expected average error multiplied by the length of the interval.

Let us restate the learning problem in the continuous limit. For simplicity we will assume that the left and the right endpoint are labeled for free (this changes the number of queries by 2).

Question 3 (Learning the continuous limit of the 1-d Ising model) *We start with an interval of length L . We are going to query B points (not counting the queries to the endpoints); for adaptive strategies the queries can depend on the answers to previous queries. For a point that lies at position x in an interval of length ℓ the answer to the query is random with distribution given by (10) and (11) (either P'_{AA} or P'_{AD} depending on whether the endpoints of the interval have the same or differing labels). Let $x_1 < \dots < x_B$ be the points queried (by convention $x_0 = 0$ and $x_{B+1} = L$) and let y_0, \dots, y_{B+1} be the corresponding labels. The expected average error is the expected value of*

$$\frac{1}{L} \sum_{i=0}^B (x_{i+1} - x_i) E'_{T_i}(x_{i+1} - x_i),$$

where $T_i = A$ if $y_i = y_{i+1}$ and $T_i = D$ otherwise. What is the optimal (adaptive) strategy that minimizes the expected average error?

The simplest non-adaptive strategy—querying uniformly spaced points is, not surprisingly, optimal (among non-adaptive strategies).

Proposition 2. *Let A be the non-adaptive strategy that queries points $x_i = i/(B+1)$ for $i = 0, \dots, B+1$. Then*

$$\text{err}_A = \frac{1}{2} \left(1 - \frac{1 - \exp(-L/(B+1))}{L/(B+1)} \right) = \frac{L}{4(B+1)} + O(1/B^2). \quad (13)$$

This is the optimal non-adaptive strategy.

Proof :

Let $\ell = L/(B+1)$ be the width of the intervals cut out by x_0, \dots, x_{B+1} . Let y_0, \dots, y_{B+1} be the labels of the points queried. We have $P(y_0 = +1) = P(y_0 = -1) = 1/2$ and (by (4)) we have

$$P(y_{i+1} = y_i) = \frac{1 + \exp(-2\ell)}{2},$$

and hence the contribution of each interval to the expected average error is

$$\begin{aligned} Z_\ell &:= \frac{(1 + e^{-2\ell})(\ell - \tanh(\ell))}{4} + \frac{(1 - e^{-2\ell})(\ell - \tanh(\ell/2))}{4} \\ &= \frac{1}{2} (\ell + e^\ell - 1). \end{aligned} \quad (14)$$

The average expected error is then

$$\begin{aligned} \text{err}_A &= \frac{B+1}{L} Z_\ell = \frac{1}{2} \left(1 - \frac{1 - \exp(-L/(B+1))}{L/(B+1)} \right) \\ &\leq \min \left\{ 1/2, \frac{L}{4(B+1)} \right\}, \end{aligned} \quad (15)$$

where the last inequality follows from $\exp(-\ell) \leq 1 - \ell + \ell^2/2$ for $\ell \geq 0$.

Note that Z_ℓ is convex (since $\exp(\ell)$ is convex). Thus the selection of x_1, \dots, x_B that minimizes the expected average error is uniform. ■

The contributions of intervals of length ℓ to the (un-normalized) expected average error satisfy the following inequalities

$$\ell E'_A(\ell) = \frac{1}{2} (\ell - \tanh(\ell)) \leq \frac{\ell^3}{6}, \quad (16)$$

and

$$\frac{\ell}{4} \leq \ell E'_D(\ell) = \frac{1}{2} (\ell - \tanh(\ell/2)) \leq \frac{\ell}{2}. \quad (17)$$

Note that both $\ell E'_A(\ell)$ and $\ell E'_D(\ell)$ are convex, this follows from concavity of the hyperbolic tangent for non-negative numbers. Note that $\ell E'_A(\ell) \leq \ell E'_D(\ell)$ and hence the best possible outcome (that is, the leaf of the decision tree that contributes the least to the expected average error) for any strategy is that all query points get the same label. In this case, by convexity, the average error is minimized when the query points are uniformly spaced. This is a lower bound on any adaptive strategy. Thus we have

Proposition 3. *The expected average error of any adaptive strategy is at least*

$$\frac{1}{2} \left(1 - \frac{\tanh(L/(B+1))}{L/(B+1)} \right). \quad (18)$$

For $B \geq L-1$ the lower bound can be simplified to $(L/(B+1))^2/9$. (For $B \leq L-1$ the lower bound is a constant.)

Proof :

Equation (18) follows from the discussion preceding the statement of the proposition. The simplified version of the lower bound for $B \geq L$ follows from

$$\frac{x^2}{9} \leq \frac{1}{2} \left(1 - \frac{\tanh(x)}{x} \right),$$

which is equivalent to $(1 - x + 2x^3/9) \leq (1 + x - 2x^3/9) \exp(-2x)$, which is proved using the Taylor expansion lower bound for $\exp(-2x)$. ■

Note that when B/L is less than a constant then the lower bound of Proposition 3 and the upper bound of Proposition 2 differ by a multiplicative constant

and hence in this case adaptive strategies cannot yield asymptotic improvement over nonadaptive strategies. Hence we will assume that B is sufficiently large compared to L , more specifically

$$B \geq e \cdot L. \quad (19)$$

Note that short intervals with endpoints whose labels disagree contribute much more to the expected average error (see equations (16) and (17)). Adaptive strategies can utilize binary search to rapidly decrease the contribution of “disagree” intervals: query the mid-point and recurse on the “disagree” interval of half-length. After k queries the contribution of all the intervals created to the expected average error is

$$\frac{\ell}{2^k} E'_D \left(\frac{\ell}{2^k} \right) + \sum_{i=1}^k \frac{\ell}{2^i} E'_A \left(\frac{\ell}{2^i} \right) \leq \frac{\ell}{2^{k+1}} + \frac{\ell^3}{42}. \quad (20)$$

Now we show that combining the non-adaptive strategies with binary search yields an adaptive strategy with asymptotically better performance.

Proposition 4. *There is an adaptive strategy with error bounded by $21(L/B)^2$, assuming $L \geq 1$.*

Proof :

We split our budget B into two parts (for simplicity assume $B = 2b$), start with b uniformly placed queries and then use the remaining b queries for binary search on the “disagree” intervals. The total contribution of the “agree” intervals (this includes the contributions of the second term in (20)) is bounded from above by

$$\frac{1}{6} \left(\frac{L}{b+1} \right)^3 (b+1). \quad (21)$$

The number of “disagree” intervals is stochastically dominated by a Poisson random variable with parameter L . If there are k “disagree” intervals then each has $\lfloor b/k \rfloor$ queries available, and hence will contribute $\frac{L/2}{b+1} 2^{-\lfloor b/k \rfloor}$ to the expected average error. Thus the contribution of the “disagree” intervals is bounded from above by

$$\sum_{k=0}^{\infty} \frac{\exp(-L)L^k}{k!} k \frac{L/2}{b+1} 2^{-\lfloor b/k \rfloor} \leq \frac{L^2}{b+1} \sum_{k=0}^{\infty} \frac{\exp(-L)L^k}{k!} 2^{-b/(k+1)} \leq \frac{L^2}{b+1} \left(2^{-b/T} + \sum_{k=T}^{\infty} \frac{\exp(-L)L^k}{k!} \right), \quad (22)$$

where the last inequality is true for any choice of integer T .

Assume $L \geq 1$ and let $T = b/\ln(b/L)$, $x = b/L$, and

$$c := T/L = \frac{b/L}{\ln(b/L)} = \frac{x}{\ln x}.$$

We are now going to argue that the contribution of the “disagree” intervals is not larger than the contribution of the “agree” intervals. We do not try to understand the contribution of “disagree” intervals more precisely (as this seems to be a rather unpleasant optimization). Note that

$$2^{-b/T} = L/b. \quad (23)$$

We can use the Chernoff bound for Poisson random variables (see Proposition 5 below) to obtain

$$\sum_{k=T}^{\infty} \frac{\exp(-L)L^k}{k!} \leq \left(\frac{e^{c-1}}{c^c}\right)^L \leq \frac{2}{c \ln c} \leq \frac{4L}{b}, \quad (24)$$

where the last inequality follows from

$$c \ln c = \frac{x}{\ln x} \ln\left(\frac{x}{\ln x}\right) = x - x \frac{\ln \ln x}{\ln x} \geq \frac{x}{2} = \frac{b}{2L},$$

and the second inequality follows from

$$\frac{\ln c}{2} \leq \frac{c}{4} \leq \left(\frac{c}{e}\right)^{c-1}.$$

The expected average error is bounded by $(1/L)$ times the sum of (21) and (22) and hence (using (23) and (24)) we obtain

$$\text{err}_A \leq \frac{1}{6} \left(\frac{L}{b+1}\right)^2 + \frac{5L}{b} \cdot \frac{L}{b+1} \leq 21 \left(\frac{L}{B}\right)^2. \quad (25)$$

■

Proposition 5. (see e.g. [10], p. 97) For a random variable X with Poisson distribution with parameter L and for $t \geq 1$, $P(X \geq tL) \leq \left(\frac{e^{t-1}}{t^t}\right)^L$.

We remark that the following “high probability” version of Proposition 4 also holds.

Proposition 6. For any $\delta > 0$ and budget $B \geq 2 \max\{e^2 L, \ln(1/\delta)\}$, the adaptive strategy in Proposition 4 with probability at least $1 - \delta$ (over observed labels) has average error at most

$$220 \max \left\{ L^2, (\ln(1/\delta))^2 \right\} / B^2.$$

Proof :

As noted in Proposition 4, the number of “disagree” intervals in the continuous limit follows a Poisson distribution with parameter L . Proposition 5 implies that the number of “disagree” intervals in 6 is at most $\max\{e^2 L, \ln(1/\delta)\}$ with probability at least $1 - \delta$. We use the same notation as in the proof of Proposition 4.

The worst case (over observed labels) un-normalized expected average error of the “disagree” intervals then can be bounded (using the inequalities in (17)) by

$$\max_{k \leq e^2 L, k \leq \ln(1/\delta)} k \frac{L}{2(b+1)} 2^{-\lfloor b/k \rfloor} \leq \frac{L}{b(b+1)} (\max\{e^2 L, \ln(1/\delta)\})^2,$$

where we have used the fact that $b = \frac{B}{2} \geq \max\{e^2 L, \ln(1/\delta)\}$. Combining this with the error of the “agree” intervals (21) and normalizing, we obtain the desired bound. \blacksquare

When $L \leq 1$, Proposition 4 no longer applies, since as $L \rightarrow 0$ the expected average error of any adaptive strategy is bounded from below by a linear function of L . We will show this by arguing that for a short interval with disagreeing endpoints the expected average error can decrease by a factor of at most 2^B (compared to the expected average error with no queries) in any adaptive strategy.

Proposition 7. *The un-normalized expected average error contributed by “disagreeing intervals” $f(L, B)$ of an adaptive strategy with budget B on an interval of length L conditioned on the endpoints disagreeing satisfies*

$$f(L, B) \geq \frac{1}{2} (L/2^B - \tanh(L/2^{B+1})). \quad (26)$$

Proof :

We prove the result by induction on B . For $B = 0$ the formula agrees with (12). For $B > 0$ let $x \in [0, L]$ be the point of the first query. One of the subintervals created will have disagreeing endpoints and we have one less query left. Hence

$$f(L, B) \geq \min_{x \in [0, L]} P'_{AD}(L, x) f(L - x, B - 1) + (1 - P'_{AD}(L, x)) f(x, B - 1). \quad (27)$$

Plugging the right-hand side of (26) into the right-hand side of (27) we obtain

$$\begin{aligned} & \frac{L}{2^{B+1}} - \frac{1}{4} \left(\tanh\left(\frac{L-x}{2^B}\right) + \tanh\left(\frac{x}{2^B}\right) \right) \\ & + \frac{\sinh(L-2x)}{\sinh(L)} \left(\frac{L-2x}{2^{B+1}} - \frac{1}{4} \tanh\left(\frac{L-x}{2^B}\right) + \frac{1}{4} \tanh\left(\frac{x}{2^B}\right) \right). \end{aligned} \quad (28)$$

From concavity of the hyperbolic tangent the maximum of $\tanh\left(\frac{L-x}{2^B}\right) + \tanh\left(\frac{x}{2^B}\right)$ happens for $x = L/2$. By symmetry we can assume $x \leq L/2$ (formula (28) is invariant under $x \mapsto L - x$). Since the derivative of hyperbolic tangent is in $[0, 1]$ we have

$$\frac{L-2x}{2^{B+1}} - \frac{1}{4} \tanh\left(\frac{L-x}{2^B}\right) + \frac{1}{4} \tanh\left(\frac{x}{2^B}\right) \geq \frac{L-2x}{2^{B+1}} - \frac{1}{4} \frac{L-2x}{2^B} \geq 0,$$

with the right equality holding only for $x = L/2$ (and in this case the left inequality is equality too).

Thus the minimum of (28) happens at $x = L/2$ for which (28) simplifies to

$$\frac{1}{2} (L/2^B - \tanh(L/2^{B+1}))$$

finishing the proof. ■

Proposition 8. *Assume $L \leq 1$. The expected average error of any adaptive strategy is at least*

$$\frac{L}{2^{B+5}} + \frac{L^2}{18(B+1)^2}. \quad (29)$$

Proof :

With probability $(1 - \exp(-2L))/2$ the endpoints of the interval disagree. By Proposition 7 and (17) the expected average error conditioned on the endpoints disagreeing is at least $1/2^{B+2}$ and hence the expected average error is at least

$$(1 - \exp(-2L))/2^{B+3} \geq \frac{L}{2^{B+4}}, \quad (30)$$

where the inequality uses the assumption $L \leq 1$. Equation (29) now follows by combining (30) and Proposition 3. ■

4 Proof of Proposition 1

Proof of Proposition 1:

Given a graph H consider the following problem. Each vertex v is independently assigned a uniformly random label X_v from $\{0, 1\}$. Each edge $e = \{u, v\}$ is assigned label $Y_e = X_u \oplus X_v$ (addition mod 2). Thus in total we have $|V(H)| + |E(H)|$ random variables with values in $\{0, 1\}$. Suppose we can query B of the variables and then we want to predict the other variables in such a way that we minimize the expected average error.

We are going to argue that it only makes sense to query the vertex variables and then one should choose a subset $\hat{V} \subseteq V(H)$ of them such that the subgraph of H induced by \hat{V} has the largest number of edges (which is an instance of the densest B -subgraph problem). Suppose that we query edges $\hat{E} \subseteq E(H)$ and vertices $\hat{V} \subseteq V(H)$. Let V' be the set of vertices that are connected by a path to a vertex in \hat{V} in $(V(H), \hat{E})$. Let E' be the set of edges which have both endpoints in V' or their endpoints are connected by a path in $(V(H), \hat{E})$.

Note that the values of variables in \hat{V} and \hat{E} determine the values of variables in V' and E' (for example, value of X_v for $v \in V'$ is obtained by adding the values of the edges on the path and the value of the vertex reached). Now we argue that no other variables are determined. If v is not in V' then flipping values of all vertex variables that are in the connected component of v in $(V(H), \hat{E})$ does not change value of variables in \hat{V} and \hat{E} (but it changes the value of X_v). If $e = \{u, v\}$ is not in E' then, w.l.o.g, $v \notin V'$ and by the previous sentence we

can flip values of all vertex variables that are in the connected component of v in $(V(H), \hat{E})$ (this operation changes the value of X_e but does not change the value of variables in \hat{V} and \hat{E}). Note that the value of the undetermined variables is 0 with probability 1/2 and 1 with probability 1/2.

If we have a connected component C in $(V(H), \hat{E})$ that contains a vertex in \hat{V} then instead of querying edges in $E(C)$ we can query vertices in $V(C) \setminus \hat{V}$. This will not change V' and E' and since $|V(C) \setminus \hat{V}| \leq |V(C)| - 1 \leq |E(C)|$ we did not increase the number of queries.

If we have a connected component C in $(V(H), \hat{E})$ that does not contain a vertex in \hat{V} then we can pick a vertex $v \in C$ and instead of querying edges in $E(C)$ we can query vertices in $V(C) \setminus \{v\}$. We did not increase the number of queries since $|V(C)| - 1 \leq |E(C)|$. The size of E' decreased by at most the degree of v in C and the size of V' increased by the number of vertices in $C \setminus \{v\}$. Hence the number of determined variables did not decrease.

Thus there exists an optimal strategy (that is, one that maximizes the number of determined variables) which does not query any edges. In this case E' is the set of edges induced by \hat{V} and the maximization solves the densest B -subgraph problem. ■

5 Conclusion

We have considered in this paper active prediction strategies for labelings of graphs where the labels are sampled from a graphical model. We have given a hardness result for finding the optimal set of observed nodes for general graphical models (Proposition 1). Then we have demonstrated that for the simple ferromagnetic 1-D Ising model (in the continuous limit) adaptive strategies (Proposition 4) outperform non-adaptive ones (Proposition 2). We have also demonstrated lower bounds on the prediction error (Propositions 3, 8) which match the upper bound for the $L > 1$ case. Closing the gap between the lower and upper bounds for the $L \leq 1$ case would be an interesting question. Another possible direction of future work would be to look at Ising models on more general graphs like trees and to see whether some simple adaptive selection strategy (like the one we have for 1-D Ising model) is optimal for such graphs as well.

References

1. P. Afshani, E. Chiniforooshan, R. Dorrigiv, A. Farzan, M. Mirzazadeh, N. Simjour, and H. Zarrabi-Zadeh. On the complexity of finding an unknown cut via vertex queries. In *COCOON*, pages 459–469, 2007.
2. A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *ICML*, pages 19–26, 2001.
3. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

4. N. Cesa-Bianchi, C. Gentile, and F. Vitale. Fast and optimal prediction on a labeled tree. In *COLT*, 2009.
5. N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. Active learning on trees and graphs. In *COLT*, pages 320–332, 2010.
6. U. Feige, G. Kortsarz, and D. Peleg. The dense k-subgraph problem. *Algorithmica*, 29:2001, 1999.
7. A. Guillory and J. Bilmes. Label selection on graphs. In *Conference on Neural Information Processing Systems*, 2009.
8. E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31:253–258, 1925.
9. J. M. Kleinberg and É. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. In *FOCS*, pages 14–23, 1999.
10. M. Mitzenmacher and E. Upfal. *Probability and computing*. Cambridge University Press, Cambridge, 2005. Randomized algorithms and probabilistic analysis.
11. X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.